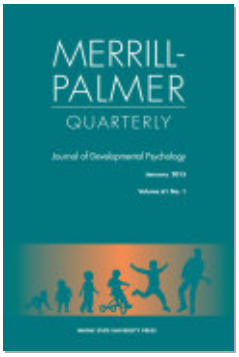
The nature of the association between moral neutralization and aggression: A systematic test of causality in early adolescence

Ribeaud, D ; Eisner, M

Abstract: This article examines possible causal linkages between moral neutralization–a generic term for the related concepts of neutralization techniques, moral disengagement, and self-serving cognitive distortions–and aggressive behavior by using a set of repeated measures in a culturally diverse urban sample at ages 11.4 and 13.7 (N = 1,032). First, correlational analyses show a strong cross-sectional association between moral neutralization and aggression. Second, fixed-effects regressions indicate substantial within-individual association implying that the cross-sectional association cannot be explained away by population heterogeneity. The within-individual association also remains stable when controlling for a number of potential confounds, which supports the notion of a direct causal relationship. Third, results of path analyses revealed near-zero lagged effects of moral neutralization on aggression when controlling for antecedent aggression and vice versa, thus suggesting no longer-term independent causal effects in either direction. Moreover, synchronous effects of moral neutralization on aggression when controlling for antecedent aggression and vice versa are same-sized and significant. Overall, results suggest a close short-term interdependence of both constructs.

DOI: https://doi.org/10.13110/merrpalmquar1982.61.1.0068

The Nature of the Association Between Moral Neutralization and Aggression: A Systematic Test of Causality in Early Adolescence



Denis Ribeaud, Manuel Eisner

Merrill-Palmer Quarterly, Volume 61, Number 1, January 2015, pp. 68-84 (Article)

Published by Wayne State University Press

The Nature of the Association Between Moral Neutralization and Aggression: A Systematic Test of Causality in Early Adolescence

Denis Ribeaud *Swiss Federal Institute of Technology, Zurich (ETH Zurich)*
Manuel Eisner *University of Cambridge*

This article examines possible causal linkages between moral neutralization–a generic term for the related concepts of neutralization techniques, moral disengagement, and self-serving cognitive distortions–and aggressive behavior by using a set of repeated measures in a culturally diverse urban sample at ages 11.4 and 13.7 ($N = 1{,}032$). First, correlational analyses show a strong cross-sectional association between moral neutralization and aggression. Second, fixed-effects regressions indicate substantial within-individual association implying that the cross-sectional association cannot be explained away by population heterogeneity. The within-individual association also remains stable when controlling for a number of potential confounds, which supports the notion of a direct causal relationship. Third, results of path analyses revealed near-zero *lagged effects* of moral neutralization on aggression when controlling for antecedent aggression and vice versa, thus suggesting no *longer-term independent* causal effects in either direction. Moreover, *synchronous effects* of moral neutralization on aggression when controlling for antecedent aggression and vice versa are same-sized and significant. Overall, results suggest a close *short-term inter*dependence of both constructs.

This article examines possible causal linkages between moral neutralization and aggressive behavior in early adolescence. *Moral neutralization* is a generic term for a set of closely related concepts from different fields of research such as the *techniques of neutralization* introduced in the

Denis Ribeaud, Chair of Sociology; Manuel Eisner, Institute of Criminology.

This study was supported by grants of the Jacobs Foundation and of the Swiss National Science Foundation.

Address correspondence to Denis Ribeaud, Criminological Research Unit, Chair of Sociology, ETH Zurich, WEP H18, Weinbergstrasse 109, CH-8092 Zurich, Switzerland. Phone: +41 44 632 97 32. Fax: +41 44 632 13 41. E-mail: ribeaudd@ethz.ch.

Merrill-Palmer Quarterly, January 2015, Vol. 61, No. 1, pp. 68–84.

field of criminology by Sykes and Matza (1957), *moral disengagement* as an element of Bandura's social cognitive theory (e.g., Bandura, Barbaranelli, & Caprara, 1996), or *secondary self-serving cognitive distortions* proposed by Gibbs (e.g., Barriga & Gibbs, 1996) in the field of developmental psychology and young-offender treatment. Ribeaud and Eisner (2010a) offer a detailed overview on the conceptual and empirical overlap among these concepts. In substance, moral neutralization refers to a set of cognitive processes through which an individual who is generally rule abiding and compliant with moral standards can minimize cognitive dissonance, threats to self-concept, and experiences of moral self-sanction when she or he transgresses those standards (Ribeaud & Eisner, 2010a). Put more simply, moral neutralization refers to (self-) justifications of moral transgressions and entails four key mechanisms: (a) cognitive restructuring or reframing of reprehensible behavior, (b) minimizing one's own agency or responsibility, (c) disregarding or distorting the negative impact of detrimental behavior, and (d) blaming, dehumanizing, or denying the victim.

In the last two decades, much evidence has been brought forth in support of a substantial association between moral neutralization and detrimental behavior in general, and aggressive behavior in particular (for overviews, see, e.g., Fritsche, 2005; Gini, Pozzoli, & Hymel, 2014; Obermann, 2011; Ribeaud & Eisner, 2010a). However, most of this research is cross-sectional (Fritsche, 2005; Maruna & Copes, 2005) and thus fails to establish the nature of the temporal and causal order between moral neutralization and aggressive behavior. We are aware of only three significant studies that analyzed the relationship longitudinally. Agnew (1994) found a small, yet significant, independent effect of prior neutralizations on later violence ($\beta = .08$) when controlling for prior violence and other possible confounds in a representative adolescent sample followed over 1 year. Paciello, Fida, Tramontano, Lupinetti, and Caprara (2008) found a correlation between trajectories of moral disengagement in adolescence and later aggression at age 20. Hyde, Shaw, and Moilanen (2010) found a substantial ($\beta = .34$) independent effect of moral disengagement at age 15 on antisocial behavior 1–2 years later when controlling for social information processing. Importantly, however, the model did not control for antecedent aggression.

Disentangling the temporal order is important because it provides key evidence on the causal direction of the link between moral neutralization and aggression and other detrimental behavior. Only if moral neutralization precedes aggression can it be conceived as a cause or at least as a facilitator of detrimental behavior. Otherwise, it would need to

be conceptualized as a consequence or cognitive reflection of detrimental behavior, in the sense of *ex post* rationalizations. In this respect, different theoretical approaches provide mixed hypotheses. According to Bandura's social cognitive theory, "people do not ordinarily engage in reprehensible conduct until they have justified to themselves the rightness of their actions" (Bandura et al., 1996, p. 365). In the same vein, and more generally, Bandura (1991) assumes that "most human behavior, being purposive, is regulated by forethought" (p. 248). Hence, in this perspective, processes of moral disengagement are explicitly conceptualized as preceding detrimental behavior and as being causally involved in its generation (for a description of the assumed causal model, see Bandura et al., 1996, pp. 366–367). Similarly, but only as a possibility, in their neutralization theory, Sykes and Matza (1957) assumed that "there is also reason to believe that [justifications for deviance] precede deviant behavior and make deviant behavior possible" (p. 666). Finally, when Barriga and Gibbs (1996) state that "secondary cognitive distortions have been characterized as pre- or post-transgression rationalizations that serve to 'neutralize' conscience or guilt and thereby to prevent damage to the self-image following antisocial behavior" (p. 334), their reception of the two previous approaches remains ambivalent with regard to the temporal order that relates the two constructs.

This theoretical ambivalence regarding the temporal order calls for empirical elucidation. Obviously, only experimental or longitudinal designs are suited to test assumptions of temporal order and to confirm or refute possible causal links that relate aggression and moral neutralization (see also Maruna & Copes, 2005, p. 45). Such contributions being scarce, the present article seeks to address this gap. Specifically, the aim of this study is to determine within a longitudinal framework to what extent moral neutralization and aggression are directly causally linked to each other and to examine the temporal order underlying such a causal relationship in early adolescence. To this end, we first assess the cross-sectional association between aggression and moral neutralization. Once this association has been established, the main focus is on a sequential test of hypotheses regarding the nature of the relationship. First, we explore whether the relationship is spurious–that is, whether it can be explained away by unobserved and by observed time-varying and time-invariant covariates that previous research has identified as key predictors of aggression and delinquency in a large range of risk domains (see, e.g., Farrington, 1998; Hawkins et al., 1998; Ribeaud & Eisner, 2010b; Wikström & Butterworth, 2006). Specifically, we consider self-control as a key personality characteristic related to aggression and

delinquency (Gottfredson & Hirschi, 1990) as well as parenting behavior, substance use, adult media use, deviant peers, unstructured leisure activities, peer victimization, parental socioeconomic status (SES), migration background, gender, and age. The focus of this part of the analysis is on *within-individual change*, which has, to our knowledge, not yet been analyzed in the research on the link between moral neutralization and aggressive behavior.

Second, if the relationship between moral neutralization and aggression remains stable when controlling for these factors (i.e., should there be evidence for a direct within-individual relationship), we then scrutinize its direction and timing. Specifically, we compare the effect sizes of aggressive behavior on later moral neutralization with those of moral neutralization on later aggressive behavior. Moreover, we examine the short-term reciprocal effects of both constructs on each other.

Within-individual change being at the core of this study, it appears judicious to focus on early adolescence, a biographical stage characterized by change and transition in many of the aforementioned risk domains (e.g., Steinberg & Silverberg, 1986).

Method

Participants

The analyses are based on data from the Zurich Project on the Social Development of Children and Youths, an experimental, prospective longitudinal study of the development of aggressive and other antisocial behavior that was set up in a culturally diverse urban context in Europe (e.g., Eisner, Ribeaud, Jünger, & Meidert, 2008; Ribeaud & Eisner, 2010b). The target sample (i.e., eligible children) consisted of all 1,675 children who entered one of 56 randomly selected public schools in Zurich, Switzerland, at age 7 in 2004. At Wave 1, 1,361 children participated, with a higher participation rate among children of German-speaking primary caregivers (91%) as compared to children of all other primary caregivers (82%). This indicates a lower participation among children with a migration background (Eisner et al., 2008, pp. 77ff.). Overall, the sample is representative of the city's child population.

At the time of the manuscript submission, five waves of data collection had been completed between the ages of 7.5 and 13.7. For the present study, we used data from child assessments in Waves 4 and 5, at ages $M = 11.4$ years and $M = 13.7$ years, respectively (henceforth referred to as "age 11" and "age 13"), when the two key measures assessed in this study–moral neutralization and the extended aggressive behavior–were

first introduced and data were collected through self-report questionnaires. Overall, 1,032 cases with complete data at both waves were available for analysis: 62% of the target sample and 76% of the wave 1 sample; 51% male. The majority of students (89%) were born in Switzerland and both biological parents of 45% were born abroad, chiefly in former Yugoslavian Republics, Sri Lanka, Germany, Portugal, and Turkey. At age 13, 71% were living with both biological parents. Compared to the sample at Wave 1, panel attrition at Waves 4/5 was significantly higher among children with a migration background than among children with at least one parent born in Switzerland (28% vs. 20%).

Procedures

Prior to data collection, parents were informed of the study in writing. In Wave 4 (age 11) parents were required to sign a consent form (active consent) in order for their child to participate; in Wave 5 parents were given the opportunity to refuse their child's participation in the study (passive consent). At the start of each data collection, participants were informed in detail about the study and about their rights, in particular the right not to answer particular questions. Participants were then asked to provide written informed consent.

Data were collected in classrooms via paper-and-pencil surveys completed in 90-minute sessions conducted in groups of 5–15 participants. Hence, all data used in this study are self-reported by participating students. Participants were guided through the questionnaire by two or three trained staff members. At 11 the data were collected during regular school lessons, whereas at age 13 data were collected during leisure time. For this reason, at age 13 participants were given a participation incentive in cash worth US$30.

Measures

Moral neutralization was measured with the 16-item instrument developed by Ribeaud and Eisner (2010a) and Ribeaud (2012), based on scales derived from the three theoretical approaches described in the introduction, including items from Bandura et al.'s (1996) moral disengagement scale, Hymel, Rocke-Henderson, and Bonanno's (2005) bullying-focused moral disengagement scale, and Huizinga and Esbensen's (1990) short neutralization techniques scale used in the Denver Youth Survey, and from a Dutch adaptation of Barriga and Gibbs's (1996) "How I Think" questionnaire that specifically focuses on self-serving cognitive distortions

related to aggressive behavior (van der Velden, 2008). The scale covers the four key mechanisms of moral neutralization: cognitive restructuring (7 items), blaming the victim (3 items), distorting negative impact (3 items), and minimizing own agency (2 items). Responses to all items were made on a 4-point Likert scale. Confirmatory factor analyses suggest an acceptably equivalent, one-dimensional factor structure across waves (Ribeaud, 2012, pp. 5f.), with high and stable reliability coefficients of $\alpha = .87$ at age 11 and $\alpha = .89$ at age 13. In the present study, a mean-score scale was used, with higher scores reflecting greater moral neutralization (Table 1).

Aggression was measured with the 12-item aggression subscale of the Social Behavior Questionnaire (Tremblay et al., 1991), adapted for adolescents, assessing physical, proactive, reactive, and indirect aggression in the last 12 months on a 5-point Likert scale. Confirmatory factor analyses suggest an acceptably equivalent one-dimensional second-order factor structure across waves, the first level representing the four subdimensions of aggression. We found stable reliability coefficients of $\alpha = .81$ at age 11 and $\alpha = .86$ at age 13. Again, a mean-score scale was derived for the present research, with higher scores indicating greater aggressive behavior.

Time-varying covariates used in the fixed-effects regression models included low self-control (10-item mean-score scale [adapted from Grasmick, Tittle, Bursik, & Arneklev, 1993], $\alpha_{\text{age 11}} = .75$, $\alpha_{\text{age 13}} = .78$), substance use (3-item variety index of tobacco, alcohol, and cannabis use in the past year; $\alpha_{\text{age 11}} = .41$, $\alpha_{\text{age 13}} = .68$), aversive parenting (5-item mean-score scale derived from the Alabama Parenting Questionnaire assessing harsh and inconsistent parenting [Shelton, Frick, & Wootton, 1996], $\alpha_{\text{age 11}} = .66$, $\alpha_{\text{age 13}} = .69$), adult media use (3-item variety index of watching adult horror, action, and other movies; $\alpha_{\text{age 11}} = .77$, $\alpha_{\text{age 13}} = .83$), deviant friends (mean-score scale across two 6-item variety indices of substance use, violence, and theft among two best friends; $\alpha_{\text{age 11}} = .67$, $\alpha_{\text{age 13}} = .83$), unstructured leisure activities (8-item mean-score scale of unstructured and unsupervised out-of-home leisure activities; $\alpha_{\text{age 11}} = .79$, $\alpha_{\text{age 13}} = .81$), and peer victimization (4-item mean-score scale of four types of peer victimization; $\alpha_{\text{age 11}} = .72$, $\alpha_{\text{age 13}} = .77$). The *time-invariant covariates* used in the fixed-effects regression models included gender (coded 1 for boys and 2 for girls), date of birth, parental SES (International Socio-economic Index of Occupational Status; Ganzeboom, De Graaf, & Treiman, 1992), parental educational achievement (10-level scale), and migration status (1 if at least one parent was born in Switzerland and 2 otherwise).

Table 1. Descriptive statistics of the variables included in the models

Variable	Age 11		Age 13		Range/coding
	M (*SD*)	Cronbach's α	*M* (*SD*)	Cronbach's α	
Moral neutralization	1.679 (0.476)	.87	2.020 (0.539)	.89	1–4, higher scores reflect higher levels of moral neutralization
Aggression	1.517 (0.427)	.81	1.776 (0.565)	.86	1–5, higher scores reflect higher levels of aggression
Low self-control	1.951 (0.466)	.75	2.203 (0.467)	.78	1–4, higher scores reflect *lower* levels of self-control
Substance use	0.034 (0.123)	.41	0.223 (0.317)	.68	0–1
Aversive parenting	1.447 (0.428)	.66	1.550 (0.470)	.69	1–4, higher scores reflect higher levels of aversive parenting
Adult media use	0.299 (0.377)	.77	0.555 (0.395)	.83	0–1
Deviant friends	0.055 (0.105)	.67	0.147 (0.209)	.83	0–1
Unstructured leisure activities	2.502 (0.881)	.79	2.926 (0.894)	.81	1–6
Peer victimization	1.780 (0.790)	.72	1.699 (0.766)	.77	1–6
Date of birth	—	—	October 21, 1997 (0.361 years)	—	June 12, 1996–February 21, 1999
Gender	—	—	1.489 (0.500)	—	1 boy and 2 girls
Migration background	—	—	1.448 (0.497)	—	1 nonmigrant and 2 migrant (both parents born abroad)
Parental education level	—	—	5.550 (3.030)	—	1–10
Parental SES (ISEI)	—	—	47.852 (19.029)	—	16–90

Note. SES = socioeconomic status; ISEI = index of occupational status.

Results

Analytical Strategy

Using a multiple-stage analytical strategy, we first describe the development of moral neutralization and aggression across the period of observation as well as their cross-sectional association at both points of measurement. Second, the causal nature of the relationship is tested within the framework of *within-individual change models.* Specifically, three two-period fixed-effects regression models (Allison, 2009, pp. 6–12) are estimated. In the first model, we assess the extent to which the baseline association found can be accounted for by unobserved population heterogeneity (Nagin & Paternoster, 2000)–that is, by time-stable, unobserved differences in the study population that affect the levels of both aggression and moral neutralization.[1] If the association between within-individual changes (difference scores) in aggression and within-individual changes in moral neutralization is significant (i.e., if the relationship is *not* attributable to population heterogeneity), in a next step *time-varying covariates* would be included in the model. That is, we test whether the within-individual covariation (i.e., change on change) can be accounted for by changes (difference scores) in other *time-variant* characteristics, such as change related to leisure activities, substance use, or media use. Finally, in a third model, selected *time-invariant predictors* (e.g., gender or SES) are included in the model to control for their *time-varying effects* (Allison, 2009, p. 10). Should these results not refute the hypothesis of a causal relationship between aggression and moral neutralization, we examine this relationship's timing and direction by means of *cross-lagged* and *synchronous reciprocal effects models.*

Descriptive Results

From age 11 to age 13, moral neutralization scores increased significantly, $t(1{,}031) = 19.5$, $p < .001$, from $M = 1.68$ ($SD = 0.48$) to $M = 2.02$ ($SD = 0.54$). Similarly, aggression scores increased, $t(1{,}031) = 15.4$, $p < .001$, from $M = 1.52$ ($SD = 0.43$) to $M = 1.78$ ($SD = 0.56$). The two constructs are also comparatively stable across time with cross-wave correlations of $r = .397$ for moral neutralization and $r = .440$ for aggression. Furthermore, there is a large cross-sectional correlation between both constructs that remains stable across time ($r = .611$ at age 11; $r = .652$ at

1. In the framework of fixed-effects regressions, unobserved differences are controlled for by using each individual as his or her own control. De facto, in the case of two-period fixed-effects models, within-individual difference scores of the dependent variable are regressed on within-individual difference scores of the independent variable (Allison, 2009, p. 14).

age 13). Finally, moral neutralization at age 11 was moderately correlated with later aggression at age 13 ($r = .305$) and vice versa ($r = .297$). All reported correlations are significant at $p < .001$.

Fixed-Effects Regressions

Having established a strong cross-sectional, *inter*individual correlation between aggression and moral neutralization, we examined whether evidence for a causal nature of this association can be found or whether, in contrast, the relationship can be accounted for by observed and unobserved covariates. As indicated, fixed-effects regression models were used for this purpose (*xi: xtreg [ ], fe* procedure in STATA 11). Since at this stage of analysis the direction of the relationship is not yet of interest, all models were calculated both with aggression as the dependent variable (left column in Table 2) and with moral neutralization as the dependent variable (right column in Table 2). First, we tested whether the association reflects unobserved differences in the study population that account for the association (population heterogeneity). This was achieved by regressing *within-individual* change scores of moral neutralization on *within-individual* change scores of aggression and vice versa (Model 1 in Table 2). The corresponding regression weights of moral neutralization, $B = 0.551$; $SE(B) = 0.025$, and aggression, $B = 0.596$, $SE(B) = 0.027$, were significant. This means that the association between aggression and moral neutralization cannot be accounted for by time-invariant effects of unobserved population heterogeneity since the association can also be observed as a change-on-change association *within* individuals.

Next, we further included a set of time-varying covariates that previous research has identified as key predictors of aggressive behavior and juvenile delinquency, the assumption being that within-individual change in these variables might account for the within-individual association found in Model 1. The coefficients of moral neutralization and aggression in Model 2 suggest that these covariates somewhat attenuate the association between aggression and moral neutralization. However, the effects are still considerable, $B = 0.427$, $SE(B) = 0.026$, and $B = 0.477$, $SE(B) = 0.029$, and significant. When further extending the model (Model 3) by allowing for *time-variant effects* of *time-invariant covariates*–including age, gender, SES, parental educational achievement, and migration status–the association between moral neutralization and aggression remained significant and almost unaffected as compared to Model 2. Hence, provided that key covariates have not been omitted, the results of the fixed-effects regression could not refute the hypothesis of a direct causal relationship between moral neutralization and aggression.

Table 2. Fixed-effects regression models, effect sizes of moral neutralization on aggression (left side) and vice versa (right side)

	Dependent: Aggression			Dependent: Moral neutralization		
	B	*SE(B)*	*p(B)*	*B*	*SE(B)*	*p(B)*
Model 1						
Moral neutralization	0.551	0.025	.000	—	—	—
Aggression	—	—	—	0.596	0.027	.000
Time dummy	0.071	0.016	.000	0.187	0.016	.000
Intercept	0.593	0.042	.000	0.775	0.042	.000
Model 2 (Model 1 + time-varying covariates)						
Moral neutralization	0.427	0.026	.000	—	—	—
Aggression	—	—	—	0.477	0.029	.000
*Low self-control	0.171	0.030	.000	0.173	0.031	.000
*Substance use	0.039	0.050	.441	0.084	0.053	.111
*Aversive parenting	0.120	0.028	.000	0.115	0.030	.000
*Adult media use	–0.007	0.035	.831	0.003	0.036	.934
*Deviant friends	0.289	0.082	.000	0.195	0.087	.025
*Unstructured leisure activities	0.013	0.015	.385	0.028	0.015	.067
*Peer victimization	0.071	0.015	.000	–0.023	0.016	.147
Time dummy	0.028	0.019	.136	0.115	0.019	.000
Intercept	0.119	0.070	.092	0.407	0.073	.000
Model 3 (Model 2 + time-invariant covariates)						
Moral neutralization	0.419	0.026	.000	—	—	—
Aggression	—	—	—	0.477	0.030	.000
*Low self-control	0.168	0.029	.000	0.171	0.031	.000
*Substance use	0.031	0.050	.538	0.094	0.053	.079
*Aversive parenting	0.132	0.028	.000	0.108	0.030	.000
*Adult media use	–0.028	0.035	.412	0.006	0.037	.861
*Deviant friends	0.289	0.082	.000	0.196	0.087	.025
*Unstructured leisure activities	0.017	0.015	.235	0.028	0.016	.069
*Peer victimization	0.065	0.015	.000	–0.023	0.016	.160

	Dependent: Aggression			Dependent: Moral neutralization		
	B	*SE(B)*	*p(B)*	*B*	*SE(B)*	*p(B)*
**Date of birth	0.000	0.000	.535	0.000	0.000	.634
**Gender	0.026	0.006	.002	0.021	0.028	.450
**Migration background	0.063	0.029	.030	–0.020	0.031	.524
**Parental education level	–0.011	0.006	.076	–0.001	0.006	.850
**SES (ISEI)	0.001	0.001	.523	–0.001	0.001	.199
Time dummy	–0.765	1.382	.580	–0.524	1.475	.722
Intercept	0.125	0.070	.075	0.420	0.074	.000

Note. *time-varying covariates, **time-invariant covariates. SES = socioeconomic status; ISEI = index of occupational status.

Cross-lagged Models (Finkel, 1995)

The next set of models focused on the direction and temporal structure of the hypothesized causal relationship between moral neutralization and aggression. In order to facilitate both model parameterization and comparisons across effect sizes, all analyses are based on *z*-standardized variables. Hence, standardized coefficient values are reported. With regard to the notation used for the model specifications and for the presentation of the results, please refer to the generic path model in Figure 1. All models were estimated with the AMOS 20 (SPSS, Chicago) structural equation modeling software by using maximum likelihood estimators.

In the baseline parameterization of the cross-lagged model, all parameters are freely estimated except β_5, and β_6, which are constrained to zero. This initial saturated model ($\chi^2 = 0$; $df = 0$) allows us to test whether

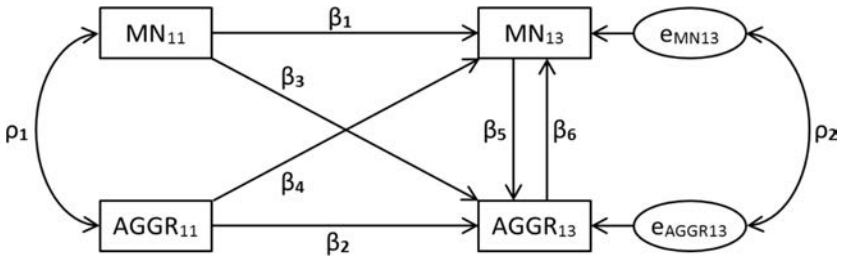


Figure 1. Generic path model of the causal relationship between moral neutralization (MN) and aggression (AGGR) from age 11 to age 13.

moral neutralization at age 11 has an independent effect on aggression at age 13 when controlling for aggression at age 11 and vice versa. With $\beta_3 = .058$ ($p = .098$) and $\beta_4 = .086$ ($p = .017$), these effects are very weak and only partially significant. Since constraining β_3 and β_4 to equality does not decrease model fit significantly ($\chi^2 = 0.218$, $df = 1$, $p = .641$), we find no evidence suggesting that the lagged effect is stronger in one direction than in the other–that is, there is no clear indication as to the main direction of the causal relationship. Finally, constraining both β_3 and β_4 to zero results in a significant decrease in model fit compared to the initial saturated model ($\chi^2 = 13.3$, $df = 2$, $p = .001$). However, the decrease in comparative fit index from 1.000 to 0.992 suggests that this restriction affects the model fit to only a very limited extent, so it appears to be an acceptable model specification. In sum, the cross-lagged effects of moral neutralization on aggression and of aggression on moral neutralization are equal and are very close to zero.

Synchronous Reciprocal Effects Models (Finkel, 1995)

Given the lack of longer-term, cross-lagged effects, we then looked at synchronous reciprocal effects. Importantly, in two-wave designs, such models can be estimated only if cross-lagged effects are (near) zero (Finkel, 1995), as is presently the case. Accordingly, in the synchronous reciprocal effects model, all parameters are freely estimated except β_3 and β_4 (i.e., the cross-lagged paths), which are constrained to zero. In substance, this model tests whether moral neutralization at age 13 has an independent effect on aggression at the same age *when controlling for aggression at age 11* and vice versa. In the initial saturated model ($\chi^2 = 0$, $df = 0$) both synchronous regression paths were (near) significant ($\beta_5 = .170$, $p = .067$; and $\beta_6 = .213$, $p = .007$). Constraining both regression weights to equality did not significantly decrease model fit ($\chi^2 = 0.218$, $df = 1$, $p = .641$). However, the significance of the parameters was increased by this constraint ($\beta_5 = \beta_6 = .194$, $p < .001$). In essence, there appears to be a significant synchronous reciprocal effect of the same size in either direction. Again, there is no evidence for a clear causal direction between aggression and moral neutralization. Eventually, we also tested a model without correlated errors–that is, $\rho_2 = 0$ ($\chi^2 = 7.168$, $df = 1$; $p = .007$)–which results in stronger reciprocal effects ($\beta_5 = .339$, $p < .001$; and $\beta_6 = .338$, $p < .001$). However, in this model, too, constraining the reciprocal effects to equality hardly affected the model fit ($\Delta\chi^2 = 0.000$, $df = 1$, $p = .995$). In sum, the synchronous effects models suggest substantial and equal effects in either direction.

Discussion

In this study, we examined the nature of the association between moral neutralization and aggression based on self-report data. In particular, we examined to what extent this association can be understood as causal in nature, as well as the timing and direction of this relationship.

First, we found a pronounced, stable cross-sectional, interindividual association between moral neutralization and aggression at ages 11 and 13. With correlations clearly above $r = .5$, this association turned out to be much stronger than what was found in most earlier studies. For example, in their recent meta-analysis, Gini et al. (2014) reported a mean correlation of $r = .28$ between aggression and moral disengagement. The exceptionally high correlations found in the present study are likely due to shared method variance (i.e., self-reports), to the use of highly reliable multiple-indicator scales for both moral neutralization and aggression, and also, importantly, to a moral neutralization scale that is specifically targeted at aggressive behavior.

There was also considerable within-individual change in both constructs over time, which allowed modeling the within-individual relationship between both constructs. Within-individual models of change have the advantage of controlling for population heterogeneity (i.e., for unobserved differences in the sample population that account for both moral neutralization and aggression). The corresponding fixed-effects regression models suggested that, over a period of 2 years in early adolescence, changes in moral neutralization covaried substantially with changes in aggression *within individuals*, which is much stronger evidence for a causal relationship between the two constructs than are between-individual correlations. We further explored whether the within-individual association was possibly not genuine but rather reflected other processes of within-individual change known to be associated with the development of aggressive and delinquent behavior, such as shifts in parenting behavior, changes related to life-style and leisure activities (e.g., onset of substance use, association with delinquent peers, changes with respect to the use of adult media contents), or episodes of peer victimization. Many of these within-individual processes turned out to be associated with changes in moral neutralization and/or aggression. However, the within-individual association between moral neutralization and aggression remained highly stable and significant when controlling for these potentially confounding processes, thus suggesting a *direct* causal relationship between moral neutralization and aggression. This effect remained unaltered when we controlled for time-varying effects of time-invariant variables such a gender, SES, or migration background.

Subsequently, we examined the temporal and directional order of this relationship within the framework of two-period path models. The first set of models showed near-zero *lagged effects* of moral neutralization on aggression when controlling for antecedent aggression and vice versa, suggesting that there are no substantial *longer-term independent* causal effects in either direction. Note, however, that the effect size we found ($\beta = .07$) was virtually the same as the one found by Agnew (1994). In substance, this implies that previous moral neutralization is not–or very limitedly–predictive of *shifts* in aggressive behavior just as antecedent aggressive behavior does not appear to substantially predict changes in moral neutralization in the longer term (i.e., 2 years).

The second set of path models examined *synchronous effects* of moral neutralization on aggression when controlling for antecedent aggression and vice versa. This analysis showed significant effects of the same size in either direction. Both the lack of lagged effects and the substantial and equal reciprocal synchronous effects suggest a close *short-term inter*dependence of both constructs. Note that it is a limitation of the present study, and of all similarly designed longitudinal studies, that they cannot clearly identify cause–effect sequences that occur at time intervals shorter than the time between data-collection waves. Thus, the findings suggest reciprocal causal effects, and they suggest that causal effects had a delay of less than 2 years.

Taken together, the key findings of this research, including the very substantial direct within-individual association of change in moral neutralization with change in aggression, along with the reciprocal synchronous effects, indicate that moral neutralization and aggression are intrinsically tied to each other. That is, there is not one that can be viewed as genuinely exogenous to the other as is typically implied when moral neutralization is modeled as a predictor of aggression in most extant research. In this new perspective, moral neutralization could be conceived as the cognitive and aggression as the behavioral expression of the same phenomenon. Specifically, in the process of (aggressive) decision making, moral neutralization might be envisaged as *facilitating* aggressive behavior by providing *ex ante* justifications, whereas aggressive behavior would in turn induce *ex post* legitimizations that allow a smooth integration of norm-breaking behavior into an apparently intact moral self-concept. This interpretation is in line with Matza's (1964) conception of *soft determinism*, where effect and cause are not related in a deterministic, unidirectional way to each other. Instead, a cause (e.g., moral neutralization) affects an outcome (e.g., aggression) in a way that leaves room for individual agency in the process of decision making. In turn, the outcome affects the initial cause

in a process of feedback. This conception also comes close to Bandura's general notion of *reciprocal determinism*, in which cognitive, behavioral, and environmental factors dynamically interact and influence one another bidirectionally (e.g., see Bandura, 1991). Hence, social cognitive theory already offers a framework that would allow us to integrate the findings of the present study and to extend the current unidirectional causal model of the relationship between moral disengagement and detrimental behavior proposed by Bandura (e.g., Bandura et al., 1996).

Overall, our findings suggest that future research and theory development should focus primarily on dynamic, reciprocal processes, whereas unidirectional causal models appear of limited relevance. To test such dynamic models, it will be important to go beyond the limitations of the present study in several ways. Specifically, in order to assess the generalizability of our findings, the present research would benefit from replication in samples of different ages and cultures. Moreover, longitudinal analyses with three or more data waves would enable more refined causal models of within-individual change. Also, repeated measures at much shorter intervals would further advance our understanding of the shorter-term dynamics that link moral neutralization and aggression. Finally, experimental designs, and especially designs that entail "hot" decision making, would offer a promising alternative way to understand the short-term dynamics underlying this link.

References

Agnew, R. (1994). The techniques of neutralization and violence. *Criminology, 32*(4), 555–580. doi:10.1111/j.1745-9125.1994.tb01165.x

Allison, P. D. (2009). *Fixed effects regression models.* Thousand Oaks, CA: Sage.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248–287. doi:10.1016/0749-5978(91)90022-L

Bandura, A., Barbaranelli, C., & Caprara, G. V. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology, 71*(2), 364–374. doi:10.1037/0022-3514.71.2.364

Barriga, A. A., & Gibbs, J. C. (1996). Measuring cognitive distortion in antisocial youth: Development and preliminary validation of the "How I Think" questionnaire. *Aggressive Behavior, 22*(5), 333–343. doi:10.1002/(SICI)1098-2337(1996)22:5<333::AID-AB2>3.0.CO;2-K

Eisner, M., Ribeaud, D., Jünger, R., & Meidert, U. (2008). *Frühprävention von Gewalt und Aggression: Ergebnisse des Zürcher Interventions- und Präventionsprojektes an Schulen* [Early prevention of violence and aggression: Results of the Zurich Intervention and Prevention Project in schools]. Zürich: Rüegger.

Farrington, D. P. (1998). Predictors, causes, and correlates of male youth violence. In M. Tonry & M. H. Moore (Eds.), *Crime and justice: Vol. 24. Youth violence* (pp. 421–475). Chicago: Chicago University Press.

Finkel, S. E. (1995). *Causal analysis with panel data* (Vol. 105). Thousand Oaks, CA: Sage.

Fritsche, I. (2005). Predicting deviant behavior by neutralization: Myths and findings. *Deviant Behavior, 26*(5), 483–510. doi:10.1080/01639620968489

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*(1), 1–56.

Gini, G., Pozzoli, T., & Hymel, S. (2014). Moral disengagement among children and youth: A meta-analytic review of links to aggressive behavior. *Aggressive Behavior 40*(1), 56–68. doi:10.1002/ab.21502

Gottfredson, M. R., & Hirschi, T. (1990). *A general theory of crime.* Palo Alto, CA: Stanford University Press.

Grasmick, H. G., Tittle, C. R., Bursik, R. J. J., & Arneklev, B. J. (1993). Testing the core empirical implications of Gottfredson and Hirschi's general theory of crime. *Journal of Research in Crime and Delinquency, 30*(1), 5–29. doi:10.1177/0022427893030001002

Hawkins, J. D., Herrenkohl, T., Farrington, D. P., Brewer, D., Catalano, R. F., & Harachi, T. W. (1998). A review of predictors of youth violence. In R. Loeber & D. P. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 106–146). Thousand Oaks, CA: Sage.

Huizinga, D., & Esbensen, F. (1990). *Scales and measures of the Denver Youth Survey.* Boulder: Institute of Behavioral Science, University of Colorado.

Hyde, L. W., Shaw, D. S., & Moilanen, K. L. (2010). Developmental precursors of moral disengagement and the role of moral disengagement in the development of antisocial behavior. *Journal of Abnormal Child Psychology, 38*(2), 197–209. doi:10.1007/s10802-009-9358-5

Hymel, S., Rocke-Henderson, N., & Bonanno, R. A. (2005). Moral disengagement: A framework for understanding bullying among adolescents. *Journal of Social Sciences*, Special Issue 8, 1–11.

Maruna, S., & Copes, H. (2005). What have we learned from five decades of neutralization research? In M. Tonry & M. H. Moore (Eds.), *Crime and justice: A review of research* (Vol. 32, pp. 221–320). Chicago: University of Chicago Press.

Matza, D. (1964). *Delinquency and drift.* New York: Wiley.

Nagin, D., & Paternoster, R. (2000). Population heterogeneity and state dependence: State of the evidence and directions for future research. *Journal of Quantitative Criminology, 16*(2), 117–144. doi:10.1023/A:1007502804941

Obermann, M.-L. (2011). Moral disengagement in self-reported and peer-nominated school bullying. *Aggressive Behavior, 37*(2), 133–144. doi:10.1002/ab.20378

Paciello, M., Fida, R., Tramontano, C., Lupinetti, C., & Caprara, G. V. (2008). Stability and change of moral disengagement and its impact on aggression and violence in late adolescence. *Child Development, 79*(5), 1288–1309. doi:10.1111/j.1467-8624.2008.01189.x

Ribeaud, D. (2012). A unified measure of moral neutralization: An addendum. In M. Eisner & D. Ribeaud (Eds.), *Forschungsbericht aus der Reihe z-proso: Zürcher Projekt zur sozialen Entwicklung von Kindern und Jugendlichen* [Research report series z-proso: Zurich Project on the Development of Children and Youths] (Vol. 15). Zürich: ETH Zürich.

Ribeaud, D., & Eisner, M. (2010a). Are moral disengagement, neutralization techniques, and self-serving cognitive distortions the same? Development of a unified scale of moral neutralization of aggression. *International Journal of Conflict and Violence, 4*(2), 298–315.

Ribeaud, D., & Eisner, M. (2010b). Risk factors for aggression in preadolescence: Risk domains, cumulative risk, and gender differences: Results from a prospective longitudinal study in a multiethnic urban sample. *European Journal of Criminology 7*(6), 460–498. doi:10.1177/1477370810378116

Shelton, K. K., Frick, P. J., & Wootton, J. (1996). Assessment of parenting practices in families of elementary school-age children. *Journal of Clinical Child Psychology, 25*(3), 317–329. doi:10.1207/s15374424jccp2503_8

Steinberg, L., & Silverberg, S. B. (1986). The vicissitudes of autonomy in early adolescence. *Child Development, 57*(4), 841–851. doi:10.2307/1130361

Sykes, G. M., & Matza, D. (1957). Techniques of neutralization: A theory of delinquency. *American Sociological Review, 22*(6), 664–670.

Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivée, S., & LeBlanc, M. (1991). Disruptive boys with stable and unstable high fighting behavior patterns during junior elementary school. *Journal of Abnormal Child Psychology, 19*(3), 285–300. doi:10.1007/BF00911232

van der Velden, M. (2008). *Morele domeinverschuiving en denkfouten bij kinderen in het zesde en achtste leerjaar van de basisschool* [Moral domain shift and thinking errors in sixth and eighth graders of elementary school] (master's thesis, University of Utrecht, Utrecht).

Wikström, P. O., & Butterworth, D. A. (2006). *Adolescent crime: Individual differences and lifestyles.* Cullompton, England: Willan.